Simulation Results of Probability Proportional Size Sampling for EIA's Monthly Natural Gas Production Survey

By Preston McDowney

EIA is currently considering sampling strategies for purposes of estimating total natural gas production in the United States via a sample survey, the EIA-914. This survey will be conducted on a monthly basis. The frame for the EIA-23 was chosen to be the frame for the EIA-914. The EIA-23 is an annual survey of approximately two thousand oil and natural gas producers that collects information on both reserves and production. It includes a large certainty group plus a probability proportional to size (PPS) sample of smaller operators.

This paper will discuss the results of efforts taken by SMG to simulate the results of the probability proportional to size (PPS) sampling presented to the ASA committee at the 2004 spring meeting (see Appendix A), and to provide rationale for choosing to use a cut off sample instead of a PPS sample.

SMG used a simulation study of natural gas production estimators to evaluate properties of the proposed PPS sample. A "pseudo frame" was created consisting of natural gas operators common to the 2000 EIA-23 sample and 2002 EIA-23 frame. This was done to adjust for not having the complete 2000 EIA-23 frame available. The certainty group was constant and new PPS sample of smaller operators was selected from the pseudo frame for each simulation run. The sampled data were used to estimate total natural gas production in 2002 via a variety of methods and compared to the true total from the (pseudo) 2002 population. The accuracy of the different methods was evaluated by computing summary statistics over all the simulation runs. The simulations showed that the estimation was not as accurate as expected.

For each simulation a random number generator was used to select a sample based on the PPS sampling design mentioned above. After a sample was selected, four different procedures were used to estimate regional and national totals of natural gas production. The first procedure, the PPS estimator, uses the inverse of the probability of selection in 2000 as the weight for the 2002 response. The remaining procedures used regression estimators to estimate total production. They assumed that the 2002 natural gas production is a function of $\hat{\beta}$ multiplied by year 2000 production, plus an error term, i.e., $y_i = \hat{\beta} x_i + e_i$. $\hat{\beta}$ is estimated either with ordinary least squares ($V(y_i) = \sigma^2$), or weighted least squares ($V(y_i) = \sigma^2 x_i$), providing an estimate for production in 2002 for operators not in the sample, $\hat{y}_i = \hat{\beta} x_i$ ($i = n+1,...,N$). The estimated total national production in 2002, $\hat{T}$, is then calculated as sum of 2002 production reported by the operators in the sample plus the $\hat{\beta}$ multiplied by the total 2000 production of the

operators not in the sample, i.e. $\hat{T} = \sum_{i \in s} y_i + \hat{\beta} \sum_{i \notin s} x_i$. The last procedure, the difference estimator, uses the previous observation as the estimate, thus $\hat{\beta}$ =1.

The ordinary least squares (OLS) and weighted least squares (WLS) estimators were calculated using several methods. The first method, $\hat{\beta}_1$, only used operators selected without certainty. The second method, $\hat{\beta}_{2,}$ used both the operators selected with certainty (probability =1), and those selected without certainty. Three different estimators were calculated for each of these two methods. The first estimator uses all of the operators for that method. The second estimator recalculated $\hat{\beta}$ by removing operators determined to be outliers based on their internally studentized residuals (see Appendix B). The third estimator removed both outliers and influential variables determined by DFFITS (see Appendix B).

In summary, estimates for a given sample were made using the following procedures:

Procedure 1.)PPS weighted
Procedure 2.)OLS modeled
- Method 1: $\hat{\beta}_1$ based on non-certainty group
  - i. All operators
  - ii. Outliers removed
  - iii. Outliers and influential observations removed
- Method 2: $\hat{\beta}_2$ based on certainty and non-certainty group
  - i. All operators
  - ii. Outliers removed
  - iii. Outliers and influential observations removed

Procedure 3.)WLS modeled
- Method 1: $\hat{\beta}_1$ based on non-certainty group
  - i. All operators
  - ii. Outliers removed
  - iii. Outliers and Influential observations removed
- Method 2: $\hat{\beta}_2$ based on certainty and non-certainty group
  - i. All operators
  - ii. Outliers removed
  - iii. Outliers and influential observations removed

Procedure 4.)Difference Estimator

Table 1 summarizes the results of ten thousand simulations. All of the estimators have a slightly negative bias. With the exception of the PPS estimator, all of the estimators produced percent errors, which were not accurate enough to achieve our goals. The PPS estimator has the lowest mean percent error, but also has an extremely large maximum error.

Table 1

% Error Estimates for Total Natural Gas Production in the US Without Using Outlier and/or Influential Observation Detection	Minimum	Mean	Maximum
Probability Proportional to Size Estimator	-5.695675	-0.914781	63.19546
WLS Estimator (Uncertainty only)	-5.938367	-3.368001	-0.6778877
WLS Estimator (Certainty and uncertainty)	-3.295912	-2.425018	-1.61E+00
OLS Estimator (Uncertainty only)	-6.049851	-3.690407	-1.498191
OLS Estimator (Certainty and uncertainty)	-4.454404	-3.707699	-3.02E+00
Difference Estimator (Using observation from the previous year as the basis for the estimate)	-4.744296	-4.007725	-3.299508
% Error Estimates for Total Natural Gas Production in the US Using Outlier and/or Influential Observation Detection	**Minimum**	**Mean**	**Maximum**
WLS Estimator With Outliers Removed (Uncertainty only)	-5.540075	-3.410945	-1.311728
WLS Estimator With Outliers Removed (Certainty and uncertainty)	-3.407033	-2.564085	-1.80E+00
WLS Estimator With Both Outliers and Influential Observations Removed (Uncertainty only)	-5.938367	-3.487079	-0.7405888
WLS Estimator With Both Outliers and Influential Observations Removed (Certainty and uncertainty)	-3.375638	-2.511971	-1.69E+00
OLS Estimator With Outliers Removed (Uncertainty only)	-6.023528	-2.779638	6.16E-01
OLS Estimator With Outliers Removed (Certainty and uncertainty)	-3.375638	-2.511971	-1.69E+00
OLS Estimator With Both Outliers and Influential Observations Removed (Uncertainty only)	-3.889203	-3.86565	-3.83E+00
OLS Estimator With Both Outliers and Influential Observations Removed (Certainty and uncertainty)	-4.371222	-3.622838	-2.928195

Even when good quality natural gas production data are available, the data appear to be highly variable from year to year due to declining production from existing wells, new wells that come on line and operators that buy and sell wells (in addition to other factors that complicate the data, such as mergers, splits, births, and deaths of operators). Precise estimation of natural gas production from a sample requires accurate production data from a previous time period. Because this is a dynamic industry it is critical that the historic time period be as close to the present as possible. Given available data, this argues for the use of the natural gas operators that report on the EIA-23 as the population from which to select respondents to the EIA-914. This link between the EIA-23 and the EIA-914 would make the respondents to the EIA-914 a subset of the respondents to the EIA-23 and it is expected that this link will ultimately improve both surveys.

A number of concerns were voiced about the proposal to sub-sample from the EIA-23, most importantly was the potential for a high non-response rate for the small operators. Furthermore, data from smaller operators are not as carefully edited as for the larger operators. The EIA-23 supplements survey results with production based on information from state agencies, industry sources, etc. The data collection methods of these other sources are not always consistent with those of the EIA-23. All of these factors would require additional attention from the RPD staff, making this sampling method more complicated and burdensome.

Given the aforementioned concerns, SMG and RPD propose using a cut-off sample designed to provide 90% coverage at the national level and comparable percent coverage in the areas of the Federal Gulf, Louisiana, New Mexico, Oklahoma, Texas, Wyoming and Other States (defined as all remaining states excluding Alaska) in which the operator produced natural gas during the report month.

Initial work indicates that using a 90% cut-off sample for the EIA-23 is a viable approach and is feasible using a sample of less than 350 respondents. RPD is working to determine the exact number of respondents that would be required in each area. This approach has the following advantages:

1. It reduces the expected heavy nonresponse of many smaller operators
2. It provides a simple sample selection rule that allows flexibility in using most up-to-date data from a variety of sources, thus minimizing the sample size

Both SMG and RPD are testing this proposed methodology.